METHOD AND SYSTEM FOR MONITORING ONLINE COMPUTER NETWORK BEHAVIOR AND CREATING ONLINE BEHAVIOR PROFILES

Background of the Invention

This invention relates to a system and method for collecting computer network traffic, particularly Internet traffic, in a manner that does not associate personally identifiable information with network usage data, and creating online behavior profiles that are unassociated with individual users. Specifically, the system and method of the invention will permit Internet Service Providers (ISP) and online merchants to monitor network usage and to create behavior profiles without violating customer confidentiality.

The Internet has rapidly grown into a center for conducting commerce with unprecedented efficiency and commercial advantage; however, the Internet also presents numerous new challenges to the development and execution of appropriate business models and processes. To design and implement effective marketing and business plans, companies need to gain a better understanding of the behavior and preferences of consumers while they are conducting Internet commerce.

In the current Internet world, it has become desirable for service providers and merchants to obtain specific information about Internet users for the purpose of improving the marketing of products and services, and tailoring products and services to meet the requirements of specific customer types. In order to obtain the most effective data describing Internet consumer behavior and preferences, it is desirable to aggregate usage data from

companies that provide Internet access to their employees, and from ISPs that provide Internet access to subscribers.

However, the collection of Internet transaction data raises many concerns about consumer confidentiality and privacy. First, participating companies and ISPs desire to maintain the confidentiality of their business information such as the number of subscribers, the geographical locations of each subscriber, and general usage data.

Additionally, many users are averse to having their actions monitored and tracked. Security concerns about the Internet have prevented many users from completing online transactions. Other users have completely stayed away from the Internet because of fears that their private information might become available to third parties in an uncontrolled manner.

Therefore, it is desirable to obtain detailed information about the behavior of users while ensuring subscriber, employee, and company privacy.

Today, there are several major approaches to collecting Internet transaction data. The first is through traditional polling techniques. In this method, user behavior profiles are developed from users' answers to questionnaires regarding their Internet use. Unfortunately, this technique suffers from bias and fails to provide the detail that marketers need.

The next approach to collecting network transaction data is by using logfiles generated by network devices such as Web servers and proxies. Logfiles provide increased detail and accuracy compared to polling techniques; however, they fail to protect user privacy and confidentiality. Logfiles generally contain a username or an Internet Protocol

(IP) address that can be used to tie behavior to a particular individual. Additionally, Web server logfiles alone are ineffective in characterizing user behavior because they only contain the cross-section Internet traffic going to that Web server; the Web server logfiles are unable to accurately capture the behavior of a consumer who accesses multiple Web sites to assist in making purchasing decisions.

The last general approach to collecting network transaction data involves the use of unique identifiers called "cookies" inserted into an Internet browser. When the user accesses a Web site on the Internet, the Web server can read the inserted cookie to obtain the unique identifier and then store details about the current transaction associated with the unique identifier. This method fails to capture Internet usage for users that have cookies disabled on their browsers and also fails to capture Internet usage on Web sites that do not participate in capturing and aggregating usage data. Since the captured data is not complete, any behavior profile created using the data cannot be representative of Internet usage in the aggregate.

Under current Federal Communications Commission (FCC) regulations, companies may have to provide protection of customer proprietary network information. By monitoring and recording detailed network information about individuals using logfiles or cookies, companies may be in violation of these FCC regulations. To date, there has been no effective way of obtaining online customer behavior profiles to allow service providers and merchants to tailor products and services better without possibly violating government regulations.

It becomes desirable, therefore, to provide a method and system where such information can be obtained while still maintaining the confidentiality of the customer (e.g., by characterizing such data in a manner free of personally identifiable information).

Summary of the Invention

In accordance with the invention, a method is provided for collecting network usage data and creating user behavior profiles therefrom. The method includes obtaining an identifier representing one or more users of a computer network, creating an anonymized identifier (AID)--defined as an identifier stripped of all personally identifiable information--using the obtained identifier, and collecting data being transmitted across the computer network. If the collected data is sent to or from a user with an anonymized identifier, a transaction record is created associating the anonymized identifier with the collected data. The record is then stored in a database.

In additional embodiments of the present invention, individual users connect using any other access media available. For example, users may connect to an ISP or intranet using broadband technology such as Integrated Services Digital Network (ISDN), Digital Subscriber Line (DSL), cable modems, fiber optic networks, satellite networks, or wireless networks.

In a yet still further aspect of the invention, each user's identifier is converted to an anonymized identifier using an encryption technique such as a one-way hashing function. In more specific embodiments, the one-way hashing function is one of the following: Secure

Hashing Algorithm 1 (SHA-1), Message Digest 4 (MD4), Message Digest 5 (MD5), or the Digital Encryption Standard (DES). User profiles are then created using the anonymized identifier for each user.

In a more specific embodiment of the present invention, collected network transaction data is matched to a particular user by monitoring packets to and from an authentication server such as a RADIUS server. Also, in a more specific embodiment of the present invention, anonymized identifiers are classified according to job function, access media, geographical location, or phone number of the user.

Also, in accordance with the present invention, a system is provided for collecting network usage data without associating personally identifiable information with such data. The system includes a communication port coupled to a computer network, where the communication port provides access to one or more servers; one or more processors; and a computer memory. The computer memory contains instructions to identify a user of a computer network; create an anonymized identifier representing the identified user; and store network transaction data associated with an anonymized identifier.

Brief Description of the Drawings

Having thus briefly described the invention, the same will become better understood from the following detailed discussion, taken in conjunction with the drawings when:

Figure 1 is a general system schematic diagram showing users connected to a Point-Of-Presence (POP) Internet Service Provider, which is in turn connected to the Internet, and then illustrated connected typically to an ISP which connects to a Web server;

Figure 2A is a schematic diagram illustrating how encryption is used to take a user's ID and create an Anonymized Identifier (AID) for purposes of tracking the session record in a transaction database;

Figure 2B is a schematic diagram illustrating a two-pass encryption method for taking a user ID and creating an anonymized identifier for tracking user sessions;

Figure 3 is a block diagram of a typical data packet illustrating how data is extracted to determine interactions by the user to the host and the number of page hits established which can be tracked in accordance with the invention; and

Figure 4 is a block diagram of a typical method for collecting network transaction data whereby a system receives a network packet, extracts information from that packet, and stores the resulting information in a database;

Figure 5 is a general schematic diagram showing a configuration of a plurality of collection engines coupled to the Internet and an aggregation server coupled to the Internet whereby the aggregation server can collect and aggregate information stored on the various collection engines;

Figure 6 shows a typical aggregation server data table containing data collected from various collection engines;

Figure 7 shows the Hypertext Transfer Protocol-specific (HTTP) fields stored in a typical aggregation server data table; and

Figure 8 shows an additional embodiment of the present invention where the functionality of the aggregation server is spread over multiple servers to increase the performance and scalability of the overall system.

Detailed Discussion of the Invention

The first embodiment of the present invention provides a system and method for collecting network transaction data without associating personally identifiable information with such data. According to this embodiment, users 101 log on to an ISP 102 in the conventional manner in order to access the Internet 104 as shown in Figure 1. Once connected, a user 101 can use a network browser such as Microsoft™ Internet Explorer™ or Netscape™ Communicator™ to access Web servers 105 on the Internet 104. Users 101 can also use any other network application to access additional network services.

According to an embodiment of the present invention, a collection engine 103 is coupled to the ISP 102 in such a manner that the collection engine 103 can monitor packets sent between users 101 and the Internet 104. The collection engine 103 is a passive device that monitors network traffic, collecting data about network transactions and recording them in a database.

In order for the collection engine 103 to create online behavioral profiles that are unassociated with individual users, the present invention uses an anonymized identifier to

represent an individual user. In this embodiment of the present invention, the anonymized identifier is preferably obtained from the username of the individual user. If usernames are unavailable, the system can use any other unique identifier (e.g., MAC address, Internet Protocol (IP) address, or wireless Mobile Subscriber ISDN (MSISDN) identifier). To maintain user anonymity, it is imperative that the original username cannot be obtained from the anonymized identifier. The present embodiment applies a one-way hashing function to the login usernames. One-way hashing functions, such as Message Digest 4 (MD4), Message Digest 5 (MD5), Secure Hashing Algorithm 1 (SHA-1), etc., are commonly used in cryptography applications including digital signatures.

Figure 2A shows an example of a unique identifier 203 being created from a username 201 and a key 204 using a one-way hashing function 202. In this example, the one-way hashing function is the Secure Hashing Algorithm (SHA) developed by the National Institute of Standards and Technology (NIST) and published as a Federal Information Processing Standard (FIPS PUB 180). The key 204 is appended to the username 201. One-way hashing function 202 is applied to the combined key 204 and username 201 to produce the anonymized identifier 203. Use of the key 204 makes it more difficult to decrypt the anonymized identifier and using a unique key for each ISP ensures usernames or other identifiers are unique across ISPs. One of skill in the art will readily appreciate that any other one-way hashing algorithm can be used with the present invention.

Figure 2B shows a two-pass method for creating online behavioral profiles that are unassociated with individual users. This two-pass method is similar to the one-pass method

shown in Figure 2A. In this embodiment, a first anonymized identifier is creating as discussed above with regard to Figure 2A. Then, the first anonymized identifier encrypted using one-way hashing function 205 along with key B 206 to create a second anonymized identifier 207. The two-pass technique allows a third party to assist without compromising the security of the resulting collected data.

When a user logged on to an ISP accesses a Web page located on a server 105, the user's workstation 101 opens a network interaction to the desired server 105 using the Internet Protocol. The network packets sent between workstation 101 and server 105 contain the network address of both devices; however, the packets do not contain a username. Thus, the collection engine 103 needs to associate a unique identifier 203 with a network IP address to record the transaction without associating any personally identifiable information with such data.

In order to create the unique identifier 203 and associate it with an IP address, the collection engine 103 needs to obtain a username. In one embodiment of the present invention, the collection engine 103 monitors the network for packets containing authentication information that associate a username with an IP address. For example, if the ISP 102 is using RADIUS to authenticate users, then the RADIUS server 107 sends an authentication packet containing a username associated with an IP address whenever a user successfully logs on to the network.

In alternative embodiments of the present invention, other authentication mechanisms may be used. In most cases, the user identifier and IP address are sent across the network

unencrypted and can be obtained by the collection engine 103; however, some authentication mechanisms may use encryption or may not be sent across the network. In some instances, the access server is configured to suggest an IP address to the RADIUS server 107; if the address is not taken, the RADIUS server 107 sends back a packet allowing the assignment. In these cases, one of ordinary skill in the art using conventional software development techniques can develop software to obtain the user identifier/IP address correlation. Some other methods that are commonly used to assign IP addresses to users are Dynamic Host Configuration Protocol (DHCP) and Bootp.

In one embodiment of the present invention, a collection engine 103 is an Intel™-based computer running Linux™. In order to maintain a high degree of security, the operating system is hardened using conventional techniques. For example, the "inetd" daemon and other unnecessary daemons are disabled to limit the possibility that an unauthorized user could gain access to the system. The collection engine 103 also includes one or more network interface cards (NIC) that allow the operating system to send and receive information across a computer network.

In some embodiments of the present invention, Internet network traffic and authentication network traffic may be sent across different networks. In this case, the collection engine 103 can use multiple NICs to monitor packets sent across the different networks. Additionally, a site may wish to monitor user activity on multiple networks. The collection engine 103 can monitor as many sites as the situation demands and the hardware supports.

Using the network and hardware configuration discussed above, we now turn to the software implementation of the collection engine 103. In accordance with the present invention, application software is installed that has been developed in a manner that is conventional and well-known to those of ordinary skill in the art, at the POP location within an ISP.

The software includes a process that monitors packets sent across the device's network interfaces as shown in Figure 4. This embodiment of the present invention begins by waiting for a network packet to be received. When a network packet is received in block 401, relevant data is extracted from the packet in block 402. The relevant data depends on the protocol of the received packet. For example, if the packet is a RADIUS packet, the relevant data would include a user identifier, an IP address, and the time of authentication. If the packet is an HTTP packet, the system extracts the relevant header information including the size of the packet and the source and destination IP addresses, and records this information along with the date and time of the request. In addition, the system also records the requested Uniform Resource Locator (URL). For other packet types, the system extracts information including the source and destination IP addresses, the source and destination ports, the size of the packet, and the time of transmission.

In the preferred embodiment of the present invention, the collection engine 103 is aware of several standard protocols including HTTP, File Transfer Protocol (FTP), RealAudio™, RealVideo™, and Windows Media™. When network interactions are made

using one of these protocols, the collection engine 103 can collect additional information such as the name of the files requested.

One embodiment of the present invention also provides additional capabilities to track user sessions. For example, when a user is browsing a Web site, the user makes a series of separate requests to a Web server. In fact, a user may make several separate requests to a Web server in order to show a single Web page. When analyzing the behavior of a user to create a profile, it is useful to think of the related requests in terms of a single session instead of as multiple sessions. For example, when a user requests a Web page, the text of that Web page is downloaded along with each image referenced by that page. The user may then browse multiple pages within that Web site.

In one embodiment of the present invention, the collection engine 103 records the beginning of an interaction in a datastore when an initial HTTP network connection is opened. The system also records the time when that interaction was opened. Additional HTTP requests are determined to be within the same interaction until the interaction ends. In one embodiment of the present invention, interactions end after an inactivity period. In an additional embodiment of the present invention, interactions remain active for Transmission Control Protocol (TCP) connections until the connection is closed using TCP flow control mechanisms.

Once data has been collected by a collection engine 103, the data can be aggregated with data collected by other collection engines. For example, an ISP may have multiple

POPs and may use a collection engine to collect data at each one. The resulting data can then be aggregated by a central aggregation server 501.

In one embodiment of the present invention, an aggregation server 501 is connected to the Internet 104 through a conventional mechanism. Additionally, one or more collection engines 103 are connected to the Internet 104. The aggregation server 501 can access each of the collection engines 103 to configure and maintain them, as well as to receive network transaction data. As discussed above, efforts are taken to maintain the security of each collection engine. For this reason, a secure mechanism for logging on to collection engines 103 and a secure mechanism to retrieve data are desirable. One embodiment of the present invention uses the Secure Shell (SSH) to provide strong authentication. This helps prevent unauthorized access to the server. SSH also provides a mechanism for encrypting the datastreams between collection engines 103 and an aggregation server 501. One of ordinary skill in the art will appreciate that many additional forms of secure login can be used, including one-time password systems and Kerberos™.

As stated above, the aggregation server 501 performs two major tasks: (1) configuration and management of collection engines 103; and (2) aggregating data from collection engines 103.

In one embodiment of the present invention, the aggregation server 501 monitors each collection engine 103 using a protocol based on the User Datagram Protocol (UDP). Every five minutes, a collection engine 103 sends a UDP packet to the aggregation server 501 signifying that the collection engine 103 is still alive. Additionally, the UDP packet also

specifies the amount of data collected and the number of users currently using the system. In this manner, the aggregation server 501 can be alerted when a collection engine 103 crashes, loses its network connection, or stops collecting data. This permits the effective management of the collection engines 103 from a central aggregation server 501.

In alternative embodiments of the present invention, the collection engines 103 implement a Simple Network Management Protocol (SNMP) Management Information Base (MIB). The MIB includes information such as the time the collection server has been active, the amount of data stored on the server, and the number of active users and network sessions.

The aggregation server 501 also performs the additional task of collecting and aggregating data from the various collection engines 103. In one embodiment of the present invention, the data is collected at least once per day by the aggregation server 501 through a secure SSH connection as discussed above. The data is then initially validated so that corrupt packet information is removed and the data is sorted to facilitate loading into the central datastore.

In some embodiments of the present invention, the collection engines do not have enough storage to permit one collection every 24 hours. In these cases, the aggregation server can collect data from the collection engine more often than every 24 hours. In one embodiment of the present invention, the UDP-based management protocol discussed above can be used to determine when a collection needs to be scheduled. In addition to the information discussed above, the UDP-based management protocol also includes the percentage of collection storage that has been used. A threshold can be set to initiate a

collection. For example, if a collection engine 103 sends a UDP-based management protocol packet stating that it has used 70% of its storage capacity, then the aggregation server can initiate the process of aggregating the data from that collection engine as discussed above.

In one embodiment of the present invention, aggregation server 501 is a Sun Microsystems Enterprise 6500™ server with sixteen (16) Sparc Ultra II™ processors and four (4) Fiber Channel connections to an EMC™ disk array. The aggregation server 501 includes an Oracle™ database that is configured to store data retrieved from the various collection engines 103.

In one embodiment of the present invention, the aggregation server 501 stores the following information that is retrieved from the various collection engines 103: (1) *ISP*, a representation of an ISP that collects data; (2) *POP*, a representation for a particular point of presence within an ISP; (3) *AID*, an anonymized user identifier; (4) *Start Date*, the date and time that an interaction began; (5) *End Date*, the date and time that an interaction ended; (6) *Remote IP*, the IP address of remote host (e.g., the IP address of a Web server being accessed by a user); (7) *Remote Port*, the port of the remote computer that is being accessed; (8) *Packets To*, the total number of packets sent during the interaction; (9) *Bytes To*, the total number of bytes sent to the remote server during an interaction; (10) *Packets From*, the total number of packets received from the remote computer; (11) *Bytes From*, the total number of bytes received from the remote computer; and (12) *IP Protocol*, the protocol code used during the interaction. For example, Figure 6 shows a typical data table for the aggregation server.

Protocols such as HTTP and FTP contain additional information that can be useful in describing user behavior. One embodiment of the present invention collects additional information for these protocols. For example, Figure 7 shows a representative data table containing additional HTTP information as follows: (1) *HTTP Host*, the hostname sent as part of the HTTP request; (2) *HTTP URL*, the Uniform Resource Locator requested; (3) *HTTP Version*, the HTTP version sent as part of the request.

The various embodiments of the present invention discussed above maintain the anonymity of the user by creating and using an anonymized identifier; however, the URL used in an HTTP request may contain identifying data. One embodiment of the present invention attempts to strip identifying data from URLs before storing them. According to this embodiment, the system searches for the following words within a URL: "SID", "username", "login", and "password". If these are found, the system strips the associated identifying information. For example, if the URL were "/cgibin/shop.exe/?username=bob", then the system would strip "bob" from the URL so that this identifying information would not be stored in the aggregated database.

In one embodiment of the present invention, the aggregation server includes database-associating anonymized identifiers with a classification. For example, in one embodiment, the classification is the physical location of the user. This information is determined using the billing address of the user. There are commercial applications available that will translate a well-formed address into a Census block group code identifying the general location of that address.

In another embodiment, user classification is based on the phone number that the user dials from as transmitted using the Automatic Number Identifier (ANI) information transmitted through the Public Switched Telephone Network (PSTN), the same information used to provide the Caller ID™ service. Since the area code and exchanges of telephone numbers generally relate to a geographical area, this can be used to help identify the locality of users. For example, if ANI transmits the number 202-936-1212, the area code "202" and the exchange "936" can be used to determine the general location of the user.

In an additional embodiment of the present invention, the aggregation server 501 functionality is spread over multiple servers to increase the performance and scalability of the system as shown in Figure 8. In this embodiment, the database server 801 is a Sun Microsystems Enterprise 6500™ server as described above with reference to aggregation server 501. Database server 801 contains an Oracle™ database storing all the aggregated data.

Access server 802 is a single, secure server that gives the ability to log on to remote collection engines 103 using SSH, or some other secure mechanism, as described above. Access server 802 is the only machine that needs to have the keys necessary to securely log on to remote machines. By segregating this functionality to a single server or to a small number of servers, it is easier to monitor, configure, and maintain the access server 802 for increased security.

The access server 802 logs onto remote collection engines 103 and transfers the collected data to one load server 803. Each load server 803 receives collected data, processes

the data, and loads it into database server 801. The present invention can be embodied with one or more load servers 803. If a plurality of load servers 803 are used, any load balancing techniques can be used to distribute load across the multiple load servers 803. For example, the access server 802 can use a simple round robin technique whereby the access server 802 rotates through a list of available load servers. The access server 802 can also use a technique whereby the central processing unit load is measured for each load server 803. The server with the lowest load is given the collection to process. Other load balancing techniques are known to those of skill in the art and any such technique could be used with the present invention.

In another embodiment of the present invention, anonymized identifiers are associated with job functions. For example, a company may wish to monitor how classes of employees are using computer network resources. An anonymized identifier representing a single employee can be associated with a job function classification so that network utilization by employees with the same job function classification can be aggregated. One of ordinary skill in the art will readily appreciate that other classification systems can be used with the present invention.

Embodiments of the present invention have now been generally described in a non-limiting matter. It will be appreciated that these examples are merely illustrative of the present invention, which is defined by the following claims. Many variations and modifications will be apparent to those of ordinary skill in the art.